Clustering VoIP Caller for SPIT Identification

Muhammad Ajmal Azad*, Ricardo Morla*, Junaid Arshad† and Khaled Salah‡
*Faculty of Engineering, University of Porto, Portugal
†University of West London, United Kingdom
‡Khalifa University of Science, Technology & Research (KUSTAR),United Arab Emirates

***Abstract*—The number of unsolicited, advertisement calls (Termed as Spam over Internet Telephony (SPIT)) and messages on a typical telephony system has increased at an alarming pace. Every year, the telecommunication regulator, law enforcement agencies and telecommunication service providers receive a large number of complaints from users pertaining to these unsolicited, unwanted calls. Voice over Internet Protocol (VoIP) infrastructure is the preferred platform of choice for spammers to make these unwanted calls, primarily due to the cheap calling rate and easy integration. In this paper, we propose, design, and develop a novel detection system for unwanted calls in a telephone system, by considering the social behavior of user towards their friends, relatives, and family members. In our design, the reputation of each caller is computed by modeling the call-rate, call duration and a number of partners associated with each caller. Such information are collected from the call detailed records. Once the reputation of the caller is computed, the next step is to group the callers into spam and non-spam clusters, based on global reputation scores. The performance of the proposed approach is measured using synthetic dataset that has been generated by simulating the social behavior of spammers and non-spammers. The evaluation results show that our proposed approach is highly effective in detecting spam activities, with only 2% false positive rate under heavy and moderate SPIT attack. Additionally, the proposed approach does not require any change in the underlying VoIP network architecture, and it does not introduce any additional signalling delay.**

***Index Terms*—SPIT, Trust, Reputation, VoIP, Network Operations**

I. Introduction

Unwanted phone calls and text messages can come at any hour of the day. They annoy people at work, disturb them in their family time, and even can wake them up from sound sleep during the night. Recent statistics on the telephony spam reveal that answering spam calls can result in an estimated waste of 20 million man hours for small businesses in the United States with the annual loss of about $475 million [1]. Every year service providers, regulators, and law enforcement agencies receive thousands of complaints from consumers of the technology for these unsolicited, unauthorized, and fraudulent callers trying to abuse them. In 2012, USA FTC (Federal Trade Communication) received four times more complaints against unwanted calls than those in 2010 [2]. The number of identified spam callers has also risen to 162% from January 2013 to January 2014.

VoIP affordable calling rates, rich value added services and easy integration with the IP technologies has created an opportunity for the spammers and telemarketers to exploit the VoIP medium for unwanted, bulk un-solicited calls referred to as SPam over Internet Telephony (SPIT). Spammers normally make spam calls for advertising products, harassing subscribers, convincing subscribers to dial a premium numbers, making *Vishing*(voice equivalent of web Phishing) attack to recipient's private information, etc. Spammers can also attempt to steal user's information [3], make calls to check unsecure gateways within the network for the termination of bulk unbilled calls [4], and cause disruption in network services through flooding and denial of service attacks [5], [6].

Several approaches have been designed for combating the SPIT callers in a VoIP telephony and are mainly grouped into followings categories: a) content-based approaches that processes the speech signals on fly and blocks callers having spam content [7], [8], [9], [10], b) a list-based approaches that maintains a black, white, and grey list of callers classified as having spamming and non-spamming behavior [11],[12] c) a Turing and CAPTCHA test based approaches that ask caller to solve the challenge [13], [14], [15], [15], [16], and d) a reputation-based system that computes reputation of the caller by getting feedback from the callee or uses information from the call detail records (CDRs) [17] [18].

The installation of these existing spam detection systems in a real voice network has some concerns. The user privacy concerns, real-time processing of speech signals, encrypted speech contents, and legal issues limit the use of content processing to be used for combating spammers. The inspection of packet headers can be useful to infer the SPIT caller; however, in a VoIP network, the packet headers of legitimate and non-legitimate callers are same and do not provide any valued information to be used for classifying caller as a spammer and a non-spammer. The black and white list may control the spamming activity but has limitations as in VoIP it is easy to spoof legitimate callers identities. The complex Turing or CAPTCHA test can be difficult to be solved by the ordinary caller and spammer in a real-time voice network but it require sophisticated system and network resources for handling a large number of concurrent calls.

Mostly telephony users develop a social relationship and trust network with other users. As the user establishes more and more links (incoming and outgoing) with others, his social network grows and develops different level of trust with their interacted persons over the time. Legitimate user normally communicate within their circle of friendship and family member thus exhibits repetitive and reciprocated calling behavior. On the other hand, spammers exploit telephony for

financial gain and betraying users with the fraud thus always target large number of subscribers which results in his non-connected social network. The social network behavior of legitimate and non-legitimate callers discloses some interesting patterns that can be used for distinguishing spammer from the non-spammers. The existing social network based anti-SPIT systems mainly uses in and out-degree distribution, clustering coefficient, reciprocity etc. for blocking the spammers. However, in a voice network other features such as call duration and call rate of the caller could also provides additional information that can characterize the behavior of the caller across the network.

In this paper, we describe and evaluate a new approach that uses number of social network features for blocking SPIT caller in a VoIP network. The approach classifies caller as a spammer or non-spammers in two steps: first a global reputation is computed from caller's call rate with his callees, duration of calls with his callees and out-degree of the caller. Secondly, a dynamic threshold is computed using unsupervised machine learning approach that classifies caller as a spammers and a non-spammers. The central procedure of our approach is to create a weighted social network of the caller, compute his reputation score within the network and classify him as a SPIT or a non-SPIT. We evaluate the performance of proposed approach using synthetic data-set for different network conditions and for different percentages of spammers and non-spammer.

The paper is structured as follows: Section II briefly describes SPIT attack in a VoIP network. Section III includes review of work from the other researchers in this field. Section IV provides motivation for our approach for SPIT detection. In section V we present the features used for distinguishing SPIT caller from the non-SPIT caller and illustrate the computation of trust and reputation. The experimental setup is presented in section VI and detailed evaluation for different performance metric is presented in section VII. The paper concludes with some thoughts about the future work in section VIII.

II. Spam Over Internet Telephony

Voice spam or SPIT (Spam over Internet Telephony) are unwanted, unsolicited, pre-recorded advertisement phone calls intended to be delivered to a large number of recipients through the use of VoIP based telephony system by a spam sender that has no prior social relationship with the recipients. In the USA, during fiscal year 2012, the FTC recorded 3,840,572 consumer complaints about unwanted telemarketing calls [19] which are four times of complaints received in 2010. VoIP spammers are similar to email spammers as both have the same intent of delivering information to the recipients that contains advertisements of legal or illegal products, defraud end-users by getting private information, and spread viruses. The spam calls and messages can also be sent to and from mobile systems and the legacy PSTN telephony. The following are additional forms of spam introduced because of VoIP telephony:

Instant Message Spam: Bulk unsolicited instant messages (similar to email spam messages) but sent instantly to users of messaging system like Skype [20], WhatsApp, and Viber etc.

Presence Spam: Presence spam is bulk unsolicited set of presence requests messages to subscriber in order to get recipients buddy or white list for sending IM or call spam.

Virus Spam: Sending viruses inside bulk SMS or IM messages that affects operating system of VoIP phones and discloses system vulnerabilities to spammers.

A. SPIT Differences from E-mail Spam

SPIT exhibits some similarity with the email spam. Both email spammer and SPIT caller use Internet as medium for conveying messages but SPIT causes serious discomfort to the victims of the spam call because of real time response for the call. Besides similarity in motivations and medium, SPIT exhibits some differences from email spam with respect to spam victims and service providers. E-mail spammer utilizes text messages, images or attachments for convening their message to victims, whereas SPIT callers use digitized speech streams over Internet for conveying their messages. In terms of deciding about sender or contents, the email service provider can hold e-mails for some time period before finally delivering them to the recipient inbox, which is not noticeable to recipients. The VoIP or Voice service provider cannot hold speech stream and signalling messages without addition of noticeable delay in signalling and flow of speech streams between users. From the perspective of content processing, online processing of speech content is more challenging and resource intensive than offline processing of text messages and images.

From a user's perspective, a single e-mail spam can remain in the inbox unattended for as much time as the user wishes, but in the case of SPIT or voice call user has to respond back interactively, which makes it more annoying and disturbing. With respect to user's resource consumption, a single spam email typically consumes small number of bytes, but a voice message in a voice mail box requires greater space thus making voice mail box unavailable to the legitimate callers. In terms of human effort, the deletion of a SPIT call is more annoying and intrusive than the deletion of spam emails. In email network, the service provider assists end-users in classifying senders. Furthermore, end-users can assess legitimacy of an email by careful inspection of the subject line or the email header. On the other hand, telephony requires a user to listen to the recorded call before making a decision about its legitimacy. Moreover, there exists no system that allows service providers to provide information to callee about the nature of calls recorded in the voice mail box. Additionally, in telephony, the user might also delete some important calls if he is making decision without appropriate inspection or attention to the call. In the perspective of protocol architecture, an E-mail message is composed of two parts: header and body. The email header part can provide information about sender's nature. Telephony calls also consists of two parts: signalling and speech streams; but the signalling message though available in plain text but

is not providing any information about the sender's nature and the speech stream is only available after the call setup.

III. SPIT MITIGATION SCHEMES

Several approaches have been devised for mitigating spammers in the network. The SPIT detection systems can be grouped in two categories: content-based detection systems and identity-based spam detection systems. The content-based detection systems process the speech streams exchanged between sender and recipient using machine learning mechanism while the identity-based detection systems use identity of users (caller identity or IP-address) for monitoring the behavior of users within the network. This section provides an overview of prior works that have been carried out for countering spammers in a voice network.

A. Content-Based Approaches

SIP (Session Initiation Protocol) and RTP (Real Time Protocol) are a widely used for call setup and exchange of voice among users. The content-based approaches analyze either semantics of SIP signaling messages, or RTP contents in a real and a non-real time [7], [8], [9], [10]. The content-based approaches require advanced signal processing for speech content analysis, updated voice data-set for pattern matching. These techniques are vulnerable, when the SPIT callers intelligently change text to voice script with added noise or random text. The content-based approaches also results in a degraded voice quality, because of delay due to speech processing and also bypass user privacy. The text contents of initial call setup message of the legitimate and SPIT callers are same, thus cannot provide valuable information to distinguish a SPIT from a non-SPIT caller. The subject line of a SIP invite message may provide some information, but mostly in a real VoIP network the caller does not provide any information in the subject line, or the SPIT caller may use legitimate text in the subject line.

B. Access List-Based Approaches

Access list-based approaches compare the identities of a caller and a callee with the local and global black or white lists. A blacklist specifies who is to be kept out allowing all others to pass. A white-list only allows those who are already in the list to get through. Both of these techniques require continuous update of white and black listed users. Besides, white and black list, a gray list [11],[12] can also be used, which contains the list of callers those should be blocked on their first attempt and later allowed if a second attempt is made within a specific time window. The gray list increase the number of attempts to reach the callee. The list-based approaches are usually implemented in combination with other approaches [21], [22], [23].

C. Challenge-Response Based Approaches

The SPIT callers can be a human or a machine. The Turing tests, is a technique adopted from e-mail and depends on the fact that some things are easy for humans, but almost impossible for computers. The callers are authorized to make calls via their private-public key exchange or Turing test authentication [13], [14], [15]. The human conversation has short pause time at the beginning of an answered call followed by the statement by the callee that initiates the conversation. The overlap in a conversation patterns may consider that the caller is automated SPIT caller. These communication patterns have been analyzed in [15] and hidden Turing tests is proposed for identification of SPIT caller. A trust enforcement mechanism [16] allow caller to solve complex puzzles for generating new identities to be used for making calls using new identities. The Turing test approaches may be successful in blocking computer generated SPIT calls but it would consume more network resources. Solving puzzles increase the call setup time and put extra burden on legitimate caller to solve puzzles for every call attempt.

D. Imposing Cost on Caller

The Payments at Risk based approach [24] deducts some money from the caller account, and returns back if the caller is found legitimate at later stage. This feature is desirable because it preserves the ability of any legitimate caller to reach a large number of callee at low cost. Otherwise, users who send a large amount of wanted communication are subject to prohibitively high fees, thereby reducing the usefulness of the communication medium. The cost based approaches require an extensive micro payment infrastructure, which seems impractical.

E. Extended Call-Setup Based Approaches

Call-Setup [25] based approaches work in the following way: they accept call from the caller, disconnect it and call back caller for the call. The limitation with this technique is that it requires extra hardware or software resources and also increase the call setup time.

F. Social Reputation-Based Approaches

Social reputation based approaches use social relationship between the caller and a callee to rank caller's reputation. The trust values provide caller-callee direct trust, and the global reputation provides caller reputation as a whole in a network. The trust among users help achieving personalized detection, and global reputation is useful as a callee typically does not receive calls from all callers, and rely on the feedback from their friends or other users. In VoIP, the direct trust and global reputation can be computed in two ways: using callee feedback, or past communication pattern of the caller. The reputation of caller can be computed from the average call duration [17], the number of short and long duration calls [24], social network properties such as: node degree, local clustering coefficient, in-count, out-count, reciprocity index etc. The callee can also be asked for providing the positive or negative feedback about the callers [21], [22], [26], [27], [28].

In [17], a direct trust is computed from average call duration and global reputation is computed using Eigen trust reputation algorithm. The higher the average call duration greater the trust a callee has on a caller, and reputed the caller is. The semi-supervised clustering has been applied to callee feedback and SIP messages for clustering legitimate and non-legitimate callers [26],but it require user feedback and changes in VoIP phones. In [21], [22] a multi-stage SPIT detection system has been presented; which is based on trust and reputation of caller, and feedback about callers from other filters like black and white list. The trust is computed by getting direct feedback from the callee and the caller's reputation is computed using Bayesian inference function. The reputation base techniques can also be applied in combination with other SPIT detection approaches [24], [29], which are multistage and interact with other stages for final decision about the caller.

There are few approaches which build their filters on the basis of call duration between the caller and the callee. In [18], three system has been proposed for identifying SPIT caller in a large VoIP operator. The solutions utilize average call duration along with the page-rank algorithm for identifying possible SPIT caller. It is possible that the Spam callers make groups for SPIT attack. In [27], [28] two system has been proposed for thwarting SPIT caller by utilizing individual and group level call duration. The first system applied Mahalanobis distance to the caller call duration and time of call for distinguishing individual SPIT caller from non-SPIT callers. The second system uses entropy of the call duration at caller group level for detecting misbehaving group.

A provider level reputation system has been proposed in [30], in which the destination operator assigns reputation to caller home operator. The system enables SIP receiving operator to assign reputation score to SIP source operator by analysing the tags assigned by the source operator to the destination operator. IP Multimedia Subsystem (IMS) has been adopted for voice communication in next generation wireless network and is vulnerable to unwanted calls. In the case of IMS, the protection may be referred to as Protection against Unsolicited Communication (UC) in an (IMS) (PUCI) [31].In [32] a collaborative scorecard framework is provided for discriminating legitimate caller from the non-legitimate in an IMS network. The scores card of caller is sent to receiving domain, which built its decision whether to allow or deny caller from calling the callee in a receiving domain. The transit operator provides services for terminating VoIP traffic to specific country. In [29], a multistage SPIT detection system has been proposed and analyzed for different call rate. The system uses feedback among various stages for detecting SPIT caller in a transit VoIP operator.

G. Reputation System in Other Domains

Several approaches have been proposed for computing reputation and trust of user in other domains like P2P network, e-commerce etc. In [33], authors present a distributed trust model for computing the trust of user involved in online transactions. In [34], authors present system for computation of trust for classifying users or agents as legitimate or non-legitimate. From the perspective of email network, in [35], authors propose a method for computation of reputation of the email user based on past transactions of email user. In [36], authors presented a link prediction approach that uses the overall structure of the social graph for predicting link between two entities.

All of above mentioned systems are attempted to classify user as legitimate or non-legitimate based on connectivity network. They are only considering the feedback from the user for the other interacted users. These proposed approaches cannot be directly used for blocking spammers in a real-time VoIP medium as in VoIP and telephony, user behavior is continuously changing and can be modeled using some additional features such as duration of interaction, number of time users interacted with other, number of mutual connections etc.

IV. Discussion and Motivation

The real-time nature of VoIP requires that the SPIT callers should be refraining from calling during the call setup phase rather than examining the speech contents after the call setup. The detection of SPIT during the call setup phase not only improves customer satisfaction but also improve resource utilization. The real challenge in design of any SPIT detection system is to block a spam caller before the telephone rings without user involvement or content analysis. The content-based SPIT detection poses many challenges as it require resources for speech recognition, a speech data-set of spammers and non-spammers for real time decision, may be difficult to be applied to encrypted speech and is against the user data protection legislation. The list-based approaches need lot of maintenance and update, when calls are received from many different sources. The reputation based approaches built their filter by either getting feedback from the callee or uses average call duration, but they rely on callee for making final decision about accepting or rejecting the call.

The high average call duration is the sign of strong relationship [17], but trust between caller and the callee cannot be limited to call duration only. The SPIT caller tries to reach huge number of callee and manages to have good call duration with the large number of callees which increase his trust and global reputation as a whole. For example, consider a VoIP network having 10000 users and SPIT caller makes calls to 50% of them, who managed to have call duration of 60 seconds to 20% of called callees. In this case, although the SPIT caller has good trust with large number of callee because of good call duration which results caller having reputed behavior, however caller's structural in-balance tells story other way around. The SPIT caller can easily by-pass call-rank system [17] by having two SPIT accounts and exchange credentials with each other and by-pass the VSD system [21], [22] by giving positive feedback to SPIT identities. We believe, in addition to call duration and user feedback, other social network features can also be applied for computing trust and reputation of caller. We believe that the number of repetitive calls, number of

reciprocal calls, call duration in both direction, incoming call duration to the caller and number of unique callees to caller can provide better insight for classifying caller as spammer and non-spammer.

Our proposed approach makes twofold contribution. Firstly, a direct trust and global reputation is computed using new feature set. The direct trust and global reputation is computed from number of outgoing partners, calling rate in both direction, and total call duration in both directions. Secondly, an automated threshold is computed for global reputation, which can be used by the system for classifying caller as legitimate and non-legitimate. Our proposed work is different from other reputation based techniques in the following way: the approach uses structural features extracted from social network for computing direct trust between caller and the callee, uses power iteration method with a different initialization vector for caller's global reputation and automatic threshold for the final decision. The proposed approach can be used within the existing VoIP network without relying on users for feedback, and without changing current SIP protocol stack and network architecture.

The proposed work is different from other reputation based techniques in the following way: the approach uses various caller-callee features for computing direct trust between callers and the callee rather than average call duration only, a power iteration method with a different initialization vector for the caller's global reputation and automatic threshold for the final decision. The proposed approach can be used within the existing VoIP network without relying user for feedback, and without changing current SIP protocol stacks and network architecture.

V. PROPOSED DETECTION SCHEME

Figure below is an overview of the proposed solution for detecting SPIT caller. Caller social networks are first constructed from caller-callee past transaction from the CDR. A social network can be represented by a directed graph where caller are represented as nodes and call transaction are represented as edges. After the feature extraction and pre-processing stages, a machine learning method, such as k-means clustering, can be used for clustering the caller in two clusters. These clusters can then be analyzed for a set of features to categorize the cluster into spammer or non-spammer. The remainder of this section details the steps involved. Clustering based SPIT detection engine consists of following steps:

1) Data Processing and Computation of Caller's Direct Trust: This step involve processing of raw CDRs, present the relationship as a graph $G(V, N)$ where V is the caller S or the callee R identity and N is the trust relationship between caller and the callee. Some spammers may not be detected as spammers when considering one feature for their analysis, but may be identified as spammer using more features or combination of features. Generally, the more features one uses for direct trust computation, more likely one is to identify the non-reputed or spammer in a network. Section V-A discusses the approach for feature choices and their utility for the computation of direct trust between the caller and the callee.
2) Mapping Direct Trust into Reputation. The reputation scores are intended to give a general idea of the caller's level of participation in a system. For each selected caller we compute the reputation of caller from the caller direct trust vector using Eigen trust algorithm. The callers having small reputation values can be probably considered as non-reputed caller can be stopped from further calls. Section V-A provides the mechanism for computation of reputation.
3) Clustering Reputation of Caller: In the training process we cluster the caller having closed reputation values together. Important details of clustering algorithm affect the characteristics of cluster groups and in-turn affects the performance of detection mechanism. In section V-B we provide the mechanism for clustering the caller on basis of its reputation, how many cluster to be consider and which cluster can be consider as having suspected spit caller.
4) On-Line Classification: In the detection process, the calling behavior of caller is monitored and reputation of caller is computed using direct trust value a caller has with his interacted callees. These vectors are then compared to the constructed clustering models to measure how the observed online behavior of caller differs. In section V-C we provide the mechanism for online classification of caller.

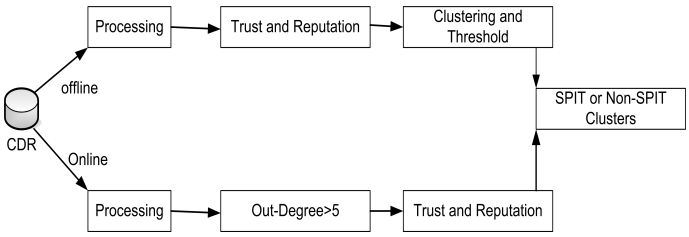

Fig. 1. SPIT Detection System

A. The Direct Trust

The direct trust presents behavior of caller towards the callee. The legitimate and non-legitimate callers have different level of direct trust towards their called callees and can be computed either by getting feedback about caller from the callee or through average call duration of caller. These features would not behave well under few situations. For example considering call duration as figure for high trust and reputation would certainly blocks legitimate caller having short duration calls and allow spammers having good duration calls. The computation of direct trust should also consider calling rate in both direction and number of unique callees a caller has.

These features are adopted from the fact that legitimate callers usually have repetitive calling nature, limited number of unique callees, have good duration with large number of called

callees and also receive good duration calls from their called callees. However the spammer called huge number of callees, which often results in a small duration calls to large a number of callees and few moderate duration calls with a few callees. In addition to this, the spammer also called a certain callee only once and hardly receives call back from the called callee. Considering all these features and observation, the direct trust between caller and callee can be computed as in equation 1.

$$Trust_{SR} = \frac{CD_{SR} \times CallRate_{SR} + CD_{RS} \times CallRate_{RS}}{PO_S} \quad (1)$$

In 1, CD is the call duration, Call-Rate is the frequency of interaction between caller and callee, and PO is the out-degree of the caller S. The normalized direct trust scores ensure that the trust scores between caller and the callee be in between 0 and 1 and can be define as equation 2.

$$T_{SR} = \frac{Trust_{SR}}{\sum_{R} Trust_{SR}} \quad (2)$$

The increase in out-partners of the caller and having small duration calls with a large number of called callee would result in a small trust value for the caller with the callee. The legitimate caller usually have small number of unique callees [37] and mostly has repetitive calls [18] with many of them so result in a long absolute call duration with many of his called callees. In addition to this the legitimate caller also receives calls from the called callees which also increase his trust toward the callee. These features makes trust of legitimate caller high as compared to non-legitimate caller having non-social network and results in a small direct trust.

B. Reputation

Once a caller's direct trust with his called callees has been computed, his reputation is to be computed for disclosing his nature across the network. It also plays an important role when the callee receives call from an unknown caller and relies on the collaboration of other network callees already communicated with the caller. The Eigen Trust algorithm is used for ranking the peers in a peer to peer network for minimizing downloads from the non-legitimate peers [38]. The detection system use power iteration method for the computation of reputation a caller across the network [17]. The better the trust caller has with large number of callees the well reputed the caller across the network and vice versa.

The power iteration algorithm is applied to the normalized trust value. The reputation of caller is computed as follows: first, for each caller S the system computes a normalized direct trust value T_{SR} with all his callee R. Secondly, a reputation score of caller S is computed by considering the direct trust scores of caller S with all his callees. This way the caller's reputation provides the wide view about his behavior towards all the callees the caller interacted across the network. The global reputation GR_S of caller S is computed as GR_S = (T_{RS})n $\times$GR$_S$; where T_{RS} is the caller-callee direct trust adjacency matrix, GR_S is a global reputation score assigned to the caller S at each iteration, and n is the number of users. The initial global reputation score of caller is set to a reciprocal of out-degree of caller S as $1/PO_S$. The initialization of global reputation to $1/PO_S$ would results in a small global reputation values to the callers having high number of unique out-going partners and high reputation scores to callers having controlled out-degree.

The caller having high number of unique callees and small call duration to a high number of callees would result in a small reputation values. These reputation values would either sent to a callee or compared with a fixed threshold for classifying caller as a legitimate or a non-legitimate.

C. Clustering and SPIT Detection

This section presents the clustering method to be used for classifying caller as a spammer and non-spammer.

D. Clustering

Clustering is an un-supervised machine learning approach needed to group the data and identifies the patterns of normal and abnormal users. The most commonly used clustering approaches are k-mean clustering and hierarchical clustering. These algorithms are based on notion of distance between the data points and use to identify data points that are close to each other.

Hierarchical Clustering compares all pairs of data points and merges the one with the closest distance. It does not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. The k-means clustering performs in an iterative process to form the number of specified clusters. The k-means method first selects a set of n points called cluster centroids as a first guess of the means of the clusters. Each observation is assigned to the nearest centroid to form a set of temporary clusters. The centroids are then replaced by the cluster means, the points are reassigned, and the process continues until no further changes occur in the clusters.

Clustering reputation data of callers faces two challenges: first finding the number of cluster that better describes the data points, second defining the threshold for declaring cluster containing spam callers. A number of methods have been proposed for finding the number of clusters that better fits the data. The With group sum of square utilizes the distance of data points from their centroid to quantify the dispersion of clusters and between groups sum of squares measure the distance of cluster centroids from the general mean of the data to quantify how apart cluster centroids are from the mean of data. The optimal solution is one having maximum BGSS and smaller WGSS.

Having performed clustering, it is necessary to distinguish the cluster that corresponds to the legitimate and non-legitimate behavior of caller. A simple and straight forward approach to filter the spammers is to consider a cluster having minimum centroid as a cluster consisting the spam caller. This can result small true positive rate when the number of

clusters increased and high false positive when the percentage of legitimate caller are higher than the non-legitimate caller. A trade-off between these values are required which work fine even when the number of clusters are increased or under conditions when different percentages of legitimate and non-legitimate traffic are present.

For detecting the anomalous caller we are using sum of the square distance between the cluster having minimum centroid values and its nearby cluster. The clusters having centroid value near to minimum centroid cluster are used for computing the threshold for classifying the caller as anomalous. The algorithm for classifying caller as anomalous and non-anomalous is presented in algorithm 2.

Algorithm 1 Detecting SPIT Caller

```
procedure SPITTER
    InPut ← Global Reputation(GR)
    OutPut ← SPIT or non-SPIT(1, −1)
    m ← kmeans(GR, k) ; k is number of clusters
    cent ← centroid(m)
    for i 1 to cent do
        if (sqrt(min(cent)-cent[i])^2)<.1 then
            x[i] ← cent
        end if
    end for
    threshold ← mean(x)
    for All caller do
        if (GR[i]<threshold) then
            Place Caller in SPIT List
        else
            Place Caller in non − SPIT List
        end if
    end for
end procedure
```

E. Classification

In classifying new caller we need to determine the window size of caller to be used for the reputation computation. At start when legitimate or non-legitimate callers are introduced in a network, they normally not to known to their friends and thus require few fixed number of calls to be made before finally comes to the reputation computation. The legitimate callers usually increases their friends slow and steady however the non-legitimate caller always tries to send as high number of calls before blocking by the operators on manually feedback from the callees. For newly introduced callers we allow him to make calls to at-least 5 unique callee before finally computing the reputation of callers. The small out-degree of caller less than 10 cannot be constitute to caller spamming behavior and is allowed even it results in a small reputation values. We use the distance of the vector from the both normal and abnormal cluster and block the caller if needed.

The deployment of this approach in a real network for spam detection poses few challenges. After computing the reputation and clustering the reputation of caller the challenge is how to baseline these clusters. The caller may exhibits different behavior in different time periods for example the caller makes more calls during the day time and makes relatively less number of calls in the night time similarly for the week days and weekend. The challenge is to determine the number of clusters as a baseline. A straight forward approach is to have a cluster for each hour for the whole week but it would increase the complexity of the system. Another approach is to have single base line for the whole day and a whole week. Though it would be less complex but it has limitation that normal behavior during weekend or peak time may corresponds to abnormal behavior during weekends or off peak times. An intermediate approach is to baseline the cluster for two time periods for week days and two time period for weekend and public holidays.

VI. Experimental Methodology

Call detailed records are almost always unavailable because of privacy reasons. However, from previous work on the analysis of call records in real telecommunication networks, we know that the call graph exhibits a power-law distribution with ($2 < \alpha < 3$) [39], [40]. In these cases, the analysed call graph exhibits a power-law distribution for in-degree with α in between 1.5 and 2, and out-degree with α in between 2 and 3 [39]. In-degree is the number of unique users calling a particular user S, and out-degree is the number of unique user a caller S is calling to. We evaluate our proposed approach using an extensive set of simulations based on: a power-law model for out-degree distribution [39], [40], the Poisson distribution for call rates [41], and the exponential distribution for the call duration [42] [43]. We repeated all simulations for at least ten times with different randomly generated network sizes to determine average values. In the following subsections we provide the description of the random data sets and the evaluation metrics used in this study.

A. Data Set

Legitimate callers exhibit power law distribution for their out-going partners with parameter $2 < \alpha < 3$ (cf. equation 3). After obtaining the out-degree distribution of the caller, the graph is created using the mechanism provided in [44].

$$p(OutPartners_S = x) = kx^{-\alpha} \tag{3}$$

Legitimate callers usually have higher calling rate within their social group, and moderate calling rate to users outside their social group. In our simulation setup the legitimate caller follows the Poisson distribution for call rate with mean $\mu = 3$ calls (cf. equation 4).

$$CallRate_{SR} = \frac{\mathrm{e}^{\mu}\mu^{\mathrm{e}}}{x!} \tag{4}$$

The legitimate caller has long duration calls with the callee within his social group, and average or short duration calls with the callee outside their social groups. In our setup, call duration exhibits an exponential distribution with average holding time $\mu = 360$ seconds using equation 5.

Network	Non-Spammer			Spammer		
	Out-Degree	In-Degree	Avg-Duration	Out-Degree	In-Degree	Avg-Duration
1% Spam	43.56054	42.71806	589.4619	941.7890	10.00634	143.1624
10% Spam	55.57226	46.88618	589.6070	953.0676	94.71745	289.1132
50% Spam	75.58833	24.36250	595.6587	954.1825	86.64567	139.2115

TABLE I
SIMULATION NETWORK STATISTICS

Prediction/Actual	Spam	Not-Spam
Spam	True Positive (A)	False Positive (B)
Not-Spam	False Negative (C)	True Negative (D)

TABLE II
CONFUSION MATRIX

$$p(CallDuration_{SR} = x) = \mu e^{-\mu x} \quad (5)$$

We also generated SPIT caller data by considering the social behavior of a SPIT caller. A SPIT caller tries to reach a large number of callees, receives a small number of incoming calls, and has short duration incoming and outgoing calls. As such, SPIT caller follows different distributions from the legitimate callers. We consider both high rate and low rate SPIT callers. The out-degree of each SPIT caller is uniformly randomly distributed between 20% and 60% of the total number of users in a network. The average call rate of SPIT caller is less than 4. The call duration of SPIT caller follows the exponential distribution with mean holding time of 180 seconds.

Finally, data generated is for 10 days consisting of 10159 users and around 1 million call records.

B. Evaluation Metric

To evaluate the performance of our system, we use standard information retrieval metrics of True Positive rate (TPR) and accuracy (ACC). The detection rate is defined as the number of spam call detected by the system divided by the total number of spam caller present in the test set. The false positive rate is defined as the total number of legitimate caller that were incorrectly classified as SPIT caller divided by total number of legitimate caller. Accuracy is the sum of true positive and false positive divided by the total number of spammer and non-spammer in a network. In order to explain these metrics, we will make use of a confusion matrix illustrated in Table 2. Each position in this matrix represents the number of elements in each original class, and how they were predicted by the classification. In Table 2, the TPR and Accuracy is computed as TPR=A/A+C and ACC=(A+D)/(A+B+C+D)

VII. PERFORMANCE EVALUATION

In this section, we evaluate the performance of propose approach for the evaluation metrics provided in above section. In addition to this we also discussed the effect of number of clusters on a detection performance and detection of SPIT caller in a real CDR.

A. Accuracy

The accurate detection system is the one which has ability of correctly classifying the nature of caller i.e. high true positive and low false positive. Unfortunately, parameter tuning usually causes these two metrics to increase or decrease at the same time. We analysed the accuracy of proposed system for different networks having different percentages of spammers and non-spammers.

We considered three networks consisting of non-spam to spam ratios of 100:100,100:10 and 100:1. The detection accuracy of our system is above 80% for all three type of networks and for any clustering parameter figure 2. However; for the network having small number of spammers, the approach achieves better accuracy with a more number of clusters. For small number of cluster, the system achieves detection accuracy of above 95% with the increase in number of spamming rate. For a small number of cluster and under low spamming rate the system achieves the accuracy of around 80%. This is due to high false positive rate and due to the fact that the few legitimate callers has small number of callees and failed to develop strong relationship with few of them. The false positive rate in this case increase with the increase in a number of clusters and get stabilizes after K=6. Comparing designed approach with call-rank, it performs better than the call-rank which achieves the accuracy rate not more than 70% for all the scenarios and cluster size.

The false positive can be minimized by introducing one more stage either in the form of vocal CAPTCHA or considering interaction history features in combination with reputation of a caller. The introduction of new stage would only affect the false positive without having any effect on true positive or other performance metric and system become more accurate.

B. Detection Rate

The true positive rate is the amount of spit caller that is detected and blocked by the SPIT detection engine. The false positive rate is the fraction of non-SPIT callers that are mistakenly considered to be SPIT by the detection engine. The VoIP operator would not tolerate both extreme that is allowing large number of SPIT caller to pass and blocking higher not number legitimate callers from calling. The operators require these ratios maximum and minimum for not losing any revenue of blocking legitimate caller. We have analysed the detection rate for two parameters that are detection rate for the number of clusters and detection rate for the amount of spammers in a network. The approach achieves a detection rate of above 90% for high and low SPIT for any number of clusters shown

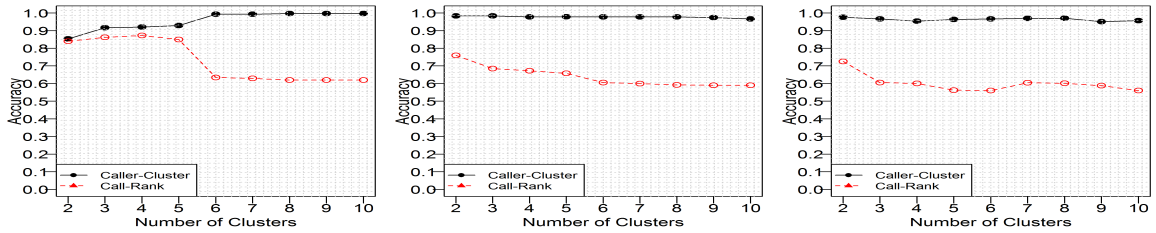

Fig. 2. SPIT Detection Accuracy for Different SPIT rate: A)SPIT Rate of 1% B)SPIT Rate of 10% C)SPIT Rate of 50%

in a figure 3 A and C. However, for the moderate rate SPIT caller, the true positive rate decreases to 80% with increase in the number of clusters shown in a figure 3.B. This decrease is because in the 10% spammer network; many spammers despite have large number of unique callees but they also have received many calls from their called callees and have average call duration greater than 300 seconds. These distributions are like the distribution of legitimate caller and remain undetected by the detection system. The true positive rate of Call-Rank decreases with the increase in number of clusters. The false positive rate for call-rank is better than the false positive rate of proposed system but still achieve false positive rate less the 5% and become stable to less the 1% with-in minimum of 6 clusters. The analyses of system for K=2 shows that the proposed system achieves true positive of 98% for any type of network compared to call-rank which achieves true positive rate of less than 90% and decreases with the increase in number of spammers.

C. Accuracy of Different Features

There are other specific social network features that can be used for identifying SPIT caller. We selected few social network features for checking their significance for detection. The detection accuracy for different social network features are presented in Table III. The ratio of spam to non-spam caller remain same as discussed above but due to space constraints cluster size is fixed to 6. The highest accuracy rate of 98% is achieved by feature communication interaction which is the ratio of unique callee the caller has, but this also decreases to 58% under high spamming attack. However this features achieves better false positive of less than 1% which is better than the false positive of our approach. But the adaptation of this feature for Spam detection may not detect the true spammers under high spam attack or where the spammers make collaborative network. We also applied entropy method to the average call duration and the caller-callee direct trust values shown in table III as entropy 1 and entropy 2. These two features achieve the detection accuracy greater than 70%. These features would not perform better alone however the combination of feature would improve the detection accuracy.

D. Addition of a New Caller

The new caller requires interaction with a number of callees so as to develop social network and has reputation values for future calling. In our approach the reputation computation

Network	Out-Degree	CI	Entropy1	Entropy2
100:100	58%	98%	93%	94%
100:10	63%	98%	82%	93%
100:1	58%	58%	70%	84%

TABLE III
ACCURACY FOR DIFFERENT FEATURES FOR K=6

requires that the caller has called at-least 5 unique callees. This threshold is because we believe a caller cannot be spammer if he calls fewer than 5 unique callees. Once the caller reached this threshold, the reputation of a caller can be computed and compared with reputation score of a spit and a non-spit cluster. The cluster having least square distance from the reputation score of a caller would be assigned to the caller. In proposed approach the reputation of a caller decreases with the increase in a number of its unique callee unless the caller has long duration and incoming calls from their called callees. Usually the newly introduced spit callers would not be able to managed high duration calls to a large number of callees and also do not receive calls from their called callee, so result in a small reputation values. These suspected callers would be blocked by the system at earliest and if missed on first attempt these would be blocked on second attempt because of decrease in a reputation values with the passage of time.

E. Complexity of Algorithm

Our design algorithm consist of three phases that are: computation of direct trust and reputation, applying k-means and detecting the anomalous caller in a network. The complexity of trust is $O(n^2)$ and reputation converges in fewer than 6 iteration for the network size of 10000 users. The complexity of the K-means clustering algorithm is O(K n t) where K is the number of clusters, n is the number of objects to be classified, and t the number of iterations which depends on the initial classification of the objects and the feature value distribution. The threshold computation and classification phase has the complexity of O(k) where k are the number of clusters. Applying the detection to new caller has O(k) complexity for detection.

VIII. CONCLUSIONS

In this paper, we have presented a novel approach for SPIT detection and filtering for VoIP networks. Our approach uses a social network features to compute the caller's trust and

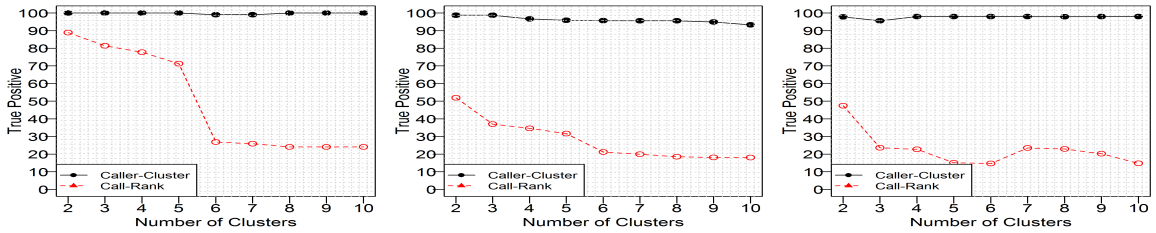

Fig. 3. Detection Rate : A)SPIT Rate of 1% B)SPIT Rate of 10% C)SPIT Rate of 50%

reputation across the network. The k-means clustering method is used for computing the threshold for distinguishing SPIT caller from the non-SPIT caller. We evaluated the system on three type of synthetic networks. Experimental measurements demonstrate that the system can achieve high accuracy for the network with high number of spammers. Overall, the true positive remains above 90% for all type of network; however, the false positive rate can be high for the network having small number of spammers. We argued and showed that this false positive rate can be improved with the addition of one more stage without affecting true positive rate. Moreover, our proposed approach and solution does not require any feedback from the caller or callee, and therefore protecting the privacy of the users. As shown, our proposed system can be deployed in a VoIP operator's network for SPIT caller identification without involving caller or the callee. In addition the system does not require any modifications to the client telephony equipment, SIP message stack or operator's network architecture.

ACKNOWLEDGMENT

The first author would like to acknowledge the financial support from the FCT (Portuguese Foundation for Science and Technology) with the associate laboratory contract INESCTEC under grant SFRH/BD/80135/2011.

REFERENCES

[1] Spam Phone Calls Cost U.S. Small Businesses Half-Billion Dollars in Lost Productivity, Marchex Study Finds. [Online]. Available: http://goo.gl/jTrgp3

[2] C. K. JENNIFER, "Complaints about Automated Calls up Sharply (last retrived August 2015)." [Online]. Available: http://goo.gl/H5HTBh

[3] M. Nassar, S. Niccolini, R. State, and T. Ewald, "Holistic VoIP Intrusion Detection and Prevention System," in *1st IPTCOMM*, 2007.

[4] R. Zhang, X. Wang, X. Yang, and X. Jiang, "Billing Attacks on SIP-based VoIP Systems," in *Ist USENIX workshop on Offensive Technologies*, 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1323276.1323280

[5] S. Ehlert, D. Geneiatakis, and T. Magedanz, "Survey Of Network Security Systems To Counter SIP-Based Denial-Of-Service Attacks," *Computers & Security*, vol. 29, no. 2, pp. 225–243, 2010.

[6] A. Keromytis, "A Comprehensive Survey of Voice over IP Security Research," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1 –24, 2011.

[7] Y. Hong, S. Kunwadee, Z. Hui, S. ZonYin, and S. Debanjan, "Incorporating Active Fingerprinting into SPIT Prevention Systems," in *The 3rd Annual VoIP Security Workshop*, 2006.

[8] D. Lentzen, G. Grutzek, H. Knospe, and C. Porschmann, "Content-Based Detection and Prevention of Spam over IP Telephony - System Design, Prototype and First Results," in *IEEEICC2011, Japan*, 2011, pp. 1–5.

[9] G. Zhang and S. Fischer-Hübner, "Detecting Near-Duplicate SPITs in voice Mailboxes using Hashes," in *14th international conference on Information security*, ser. ISC'11, 2011, pp. 152–167.

[10] A. I. Seyed, S. Hemant, and W. Haining, "A Voice Spam Filter to Clean Subscriber's Mailbox," in *8th International Conference on Security and Privacy in Communication Networks*, 2012, pp. 349–367.

[11] M. Hansen, M. Hansen, J. Mller, T. Rohwer, C. Tolkmit, and H. Waack, "Developing a Legally Compliant Reachability Management System as a Countermeasure against SPIT," in *Third Annual VoIP Security Workshop, Berlin, Germany*, 2006.

[12] D. Shin, J. Ahn, and C. Shim, "Progressive Multi Gray-Leveling: a Voice Spam Protection Algorithm," in *IEEE Network*, 2006, vol. 20, no. 5, pp. 18 –24.

[13] J. Lindqvist and M. Komu, "Cure for Spam Over Internet Telephony," in *4th IEEE CCNC*, 2007, pp. 896 –900.

[14] J. Quittek, S. Niccolini, S. Tartarelli, and R. Schlegel, "On Spam over Internet Telephony (SPIT) Prevention," in *IEEE Communications Magazine*, 2008, vol. 46, no. 8, pp. 80 –86.

[15] J. Quittek, S. Niccolini, S. Tartarelli, M. Stiemerling, M. Brunner, and T. Ewald, "Detecting SPIT Calls by Checking Human Communication Patterns," in *IEEE ICC,Scotland*, 2007, pp. 1979–1984.

[16] N. Banerjee, S. Saklikar, and S. Saha, "Anti-Vamming Trust Enforcement in Peer-to-Peer VoIP networks," in *2006 international conference on Wireless communications and mobile computing*, ser. IWCMC '06. ACM, 2006, pp. 201–206.

[17] V. Balasubramaniyan, M. Ahamad, and H. Park, "CallRank: Combating SPIT Using Call Duration, Social Networks and Global Reputation," in *Fourth CEAS2007.*, 2007.

[18] H. Bokharaei, A. Sahraei, Y. Ganjali, R. Keralapura, and A. Nucci, "You can SPIT, but You can't hide: Spammer Identification in Telephony Networks," in *2011 IEEE INFOCOM*, 2011, pp. 41–45.

[19] (2012, October) FTC Issues FY 2012 National Do Not Call Registry Data Book. [Online]. Available: http://www.ftc.gov/opa/2012/10/dncdatabook.shtm

[20] Y. Rebahi, D. Sisalem, and T. Magedanz, "SIP Spam Detection," in *ICDT '06*, 2006.

[21] P. Kolan and R. Dantu, "Socio-Technical Defense Against Voice Spamming," *ACM Trans. Auton. Adapt. Syst.*, vol. 2, no. 1, 2007.

[22] R. Dantu and P. Kolan, "Detecting Spam in VoIP Networks," in *the Steps to Reducing Unwanted Traffic on the Internet ,Berkeley, CA, USA*. USENIX, 2005, pp. 31–37.

[23] K. Ono and H. Schulzrinne, "Have I met you before?: using Cross-Media Relations to Reduce SPIT," in *3rd IPTCOMM*, 2009, pp. 1–7.

[24] Y. Rebahi, D. Sisalem, and T. Magedanz, "SIP Spam Detection," in *ICDT '06*, 2006, pp. 68–74.

[25] N. Croft and M. Olivie, "A Model for Spam Prevention in IP Telephony Networks Using Anonymous Verifying Authorities," in *ISSA 2005 new knowledge today conference, Johannesburg, South Africa*, 2005.

[26] Y.-S. Wu, S. Bagchi, N. Singh, and R. Wita, "Spam Detection in Voice-Over-IP Calls through Semi-Supervised Clustering," in *39th Annual IEEE/IFIP DSN, Portugal*, 2009, pp. 307–316.

[27] H. Sengar, X. Wang, and A. Nichols, "Thwarting Spam over Internet Telephony (SPIT) attacks on VoIP networks," in *19th IWQoS*, 2011, pp. 1–3.

[28] ——, "Call Behavioral Analysis to Thwart SPIT Attacks on VoIP Networks," in *Security and Privacy in Communication Networks*, vol. 96, 2012, pp. 501–510.

[29] M. Azad and R. Morla, "Multistage SPIT Detection in Transit VoIP," in *19 IEEE SoftCOM*, 2011, pp. 1 –9.

[30] C. Sorge and J. Seedorf, "A Provider-Level Reputation System for Assessing the Quality of SPIT Mitigation Algorithms," in *IEEE ICC '09.*, 2009, pp. 1 –6.

[31] "Study of Mechanisms for Protection against Unsolicited Communication for IMS (PUCI)," in *Release 3GPP Technical Specification*. 3GPP, 2012.
[32] A. Schmidt, A. Leicher, Y. Shah, I. Cha, and L. Guccione, "Sender Scorecards," in *IEEE Vehicular Technology Magazine*, 2011, vol. 6, no. 1, pp. 52–59.
[33] A. Abdul-Rahman and S. Hailes, "A distributed trust model," in *Proceedings of the 1997 Workshop on New Security Paradigms*, ser. NSPW '97. New York, NY, USA: ACM, 1997, pp. 48–60. [Online]. Available: http://doi.acm.org/10.1145/283699.283739
[34] G. Zacharia and P. Maes, "Trust management through reputation mechanisms,," in *Applied Artificial Intelligence ,*, 2004.
[35] J. Golbeck and J. Hendler, "Reputation network analysis for email filtering," in *IEEE conference on Email and Anti Spam*, 2004.
[36] G. Jeh and J. Widom, "Simrank: A measure of structural-context similarity," in *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 538–543. [Online]. Available: http://doi.acm.org/10.1145/775047.775126
[37] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, "Mobile call graphs: beyond power-law and lognormal distributions," in *14th ACM SIGKDD*, 2008, pp. 596–604.
[38] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The Eigentrust Algorithm for Reputation Management in P2P Networks," in *12th international conference on World Wide Web*, 2003, pp. 640–651.
[39] M. E. J. Newman, "Power laws, Pareto Distributions and Zipfs Law," in *Contemporary Physics*, 2005, vol. 46, pp. 323–351.
[40] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi, "On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications," in *15th ACM CIKM '06*, 2006, pp. 435–444.
[41] B. Chandrasekaran, "Survey of network traffic models." [Online]. Available: http://goo.gl/1dzM6B
[42] E. A. Yavuz and V. C. M. Leung, "Modeling channel occupancy times for voice traffic in cellular networks," in *2007 IEEE International Conference on Communications*, June 2007, pp. 332–337.
[43] T. Jung, S. Martin, M. Nassar, D. Ernst, and G. Leduc, "Outbound spit filter with optimal performance guarantees," *Comput. Netw.*, vol. 57, no. 7, pp. 1630–1643, May 2013.
[44] C. I. Del Genio, H. Kim, Z. Toroczkai, and K. E. Bassler, "Efficient and Exact Sampling of Simple Graphs with Given Arbitrary Degree Sequence," in *PLoS ONE*. Public Library of Science, 2010, vol. 5, no. 4, p. e10012.